"5 G" Research on System Design of Data Science and Big Data Technology Talent Training in Colleges and Universities

Xiaoying Zhang and Xin Lu(✉)

School of Science, Changchun University, Changchun 130000, China

Abstract. This paper focuses on the construction of data department and big data technology major in local colleges and Universities under the background of 5G. According to the characteristics and advantages of disciplines, it pays attention to the interdisciplinary and integration, closely follows the major needs of the country and the local, and connects the talent demand of enterprises and industries, with the guiding ideology of "through", "cross cutting", "collaboration" and "joint", It focuses on the comprehensive reform in the important links of professional development, such as training program, teaching staff construction, curriculum system construction, teaching reform and practical teaching. A set of new engineering talents training mode covering big data research and development and industrial application, reflecting the school running characteristics of local colleges and universities and the characteristics of data science and big data technology, and playing a demonstration and promotion role in local universities nationwide.

Keywords: 5G · Data science and big data technology · Local universities · Mathematics related disciplines · Talent training mode

1 Introduction

The traditional data analysis technology has been unable to meet the actual needs, and the explosive growth of data has put forward higher requirements for all walks of life. In the era background of 5G network technology, the ideological and political courses in Colleges and universities must be reasonably innovated, so as to better train the talents of the diversion. Although on the surface, the specific relationship between Ideological and political education and talent training in private colleges and Universities under the background of 5G is not obvious, and there is a suspicion that students are hard to piece together when studying on these factors [1]. However, for the relevant personnel with strong research ability, they are bound to be interested in and have the ability to complete the overall task, so as to promote the innovation of curriculum ideological and political education in private colleges and Universities under the background of 5G, To improve the quality of Ideological and political courses to implement and improve the talent training mode.

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 392, pp. 155–164, 2021.
https://doi.org/10.1007/978-3-030-87903-7_22

Under the background of the rapid development of network technology in China, new media technology has also been constantly innovated. College students do not choose to rely on traditional television, newspapers and other media to obtain information, but use the network to clarify the relevant video and audio information. Thus, the development of the network is closely related to the implementation of modern education, As the upgraded version of 4G network technology, 5G network technology can make more devices access to the network, thus forming innovative data, which has a profound impact on college education. The above is basically the change brought about by 5G network era.

2 Related Work

The downlink transmission of 5G system adopts OFDM technology, and the time-domain transmission signal of the transmitter can be expressed as:

$$s(k) = \frac{1}{N}\sum_{n=0}^{N-1} X_n e^{\frac{j2\pi n}{N}}, k = 0, 1, \ldots, N-1 \tag{1}$$

Where n is the number of FFT points and X_n is the data modulated to the nth subcarrier in frequency domain.

When the time domain signal s(k) of the transmitter passes through the multipath fading channel and the Gaussian white noise is added, the time domain complex baseband signal of the receiver can be expressed as:

$$r(k) = [\sum_{d=0}^{D-1} h_d(k)s(k - \tau d - \delta)]e^{\frac{j2\pi n}{N}} + z(k) \tag{2}$$

Where $r(k)$ is the received signal, $h_d(k)$ is the channel coefficient of the D-th path, which obeys Rayleigh distribution, D is the maximum channel delay extension, τd is the multipath delay, δ is the time offset between the sender and receiver, and $z(k)$ is the additive white Gaussian noise.

This paper analyzes the characteristics of the main synchronization sequence in 5G system. Aiming at the problem that the traditional timing synchronization algorithm can not achieve fast synchronization in 5G system with large frequency offset, an improved timing synchronization algorithm based on segment correlation is proposed [2]. The algorithm pre stores the anti frequency and frequency domain sequences locally, decomposes the long correlation into the short correlation, and transforms the data to the frequency domain to achieve fast correlation, thus effectively reducing the computational complexity.

3 Relying on the Construction of Data Science and Big Data Technology Specialty in Local Colleges and Universities

3.1 Making Training Plan and Exploring Talent Training Mode

Learning from the successful experience of the reform of the talent training program of mathematics and computing science and the establishment of "Shaofeng School of

mathematics", according to the guiding ideology of stratification + diversion and individualized development, and according to the talent training objectives, the traditional professional training mode is changed to separate undergraduate students, and the new training mode of "specialized courses + whole school supplementary courses, innovation laboratory, and school enterprise joint training" is implemented, Thus, it constructs a diversified and three-dimensional talent training mode of data department and big data technology specialty in local colleges and universities [3]. In 2018, relying on the undergraduate major of "data science and big data technology", our university applied for the construction of data science teaching experimental platform, including the purchase of software and hardware equipment and the construction of experimental environment through the central financial support for the reform and development of local colleges and universities.

3.2 Construction of Teaching Staff

In terms of optimizing the teaching staff in the University, we should break through the barriers of the University and deeply integrate the teaching staff of mathematics, statistics, computational science and other related applied disciplines. Promote the combination of big data mathematics with statistical basis, computer foundation and Application module. Build four core teams for big data collection, storage, analysis and visualization. In view of the problem that it is difficult for local universities to introduce high-level teachers in such disciplines as statistics and computer science, the university intends to give appropriate preference to policies and introduction efforts to strengthen the weak links of teachers. In terms of the construction of double qualified teachers team, centering on the "ability building as the core and system innovation as the driving force" of the construction of teaching staff, through the two-way docking of "University Research Institute (Institute) joint" and "school enterprise alliance", the construction of professional "double qualified" teachers team is carried out, and the young and middle-aged teachers with professional characteristics are guided and formed.

4 Teaching Research and Reform Simulation Analysis

Because data science and big data technology is a new major, considering its strong interdisciplinary and strong practicability, we should carry out teaching research and teaching reform in time. In terms of teaching and research, we should consider the characteristics of data science and big data technology, break the barriers between colleges and departments, build incentive measures for documents, and study the cross integration of mathematics, statistics, computer science and related application disciplines, and constantly optimize the knowledge structure of the major. Further build a complete talent training system of data science and big data technology from undergraduate, master, doctor to postdoctoral, and establish a benign interaction mechanism between discipline construction and data science and big data technology teaching. Professional teachers are encouraged to apply for teaching reform projects at provincial and ministerial level and publish high-quality teaching research papers. In the aspect of teaching reform, we should reform the teaching organization form, based on the in class teaching link, take

the extracurricular science and technology interest group, subject competition and innovative experimental plan project as the platform to carry out "on campus + off campus, in class + extra-curricular" interdisciplinary training and school enterprise joint training [4]. In September 2017, the main leader of big data undergraduate major of Xiangtan University applied for the Ministry of education's new engineering research and practice project "construction and practice of data science and big data technology in local universities" by the Ministry of education, and successfully obtained the project. Teaching research and Reform data sample is shown in Fig. 1. From Fig. 2, we can see that under the 5G background, the bonus of talent training is mainly concentrated on boys, while girls are less. Therefore, we can see that under the 5G background, both boys and girls can get more profits [5].

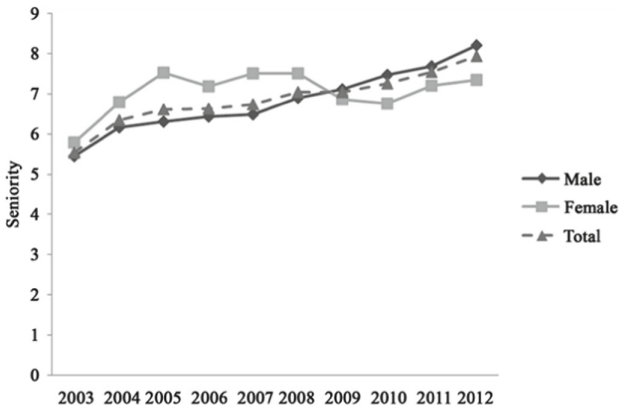

Fig. 1. Teaching research and reform

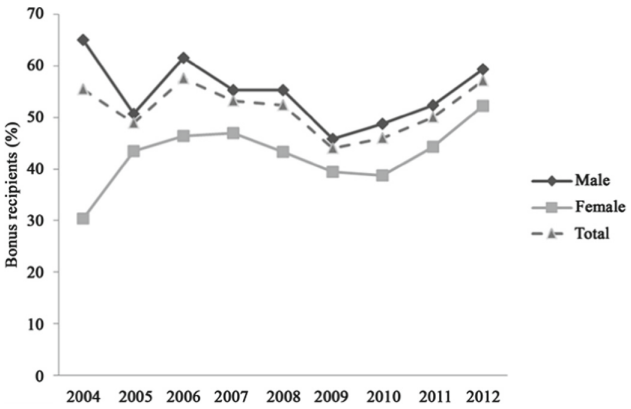

Fig. 2. Talent training bonus

5 Strengthening Practical Teaching

Because data science and big data technology are practical majors, it is necessary to have a good practical base as training support [6]. Take Xiangtan University as an example, Make full use of the Key Laboratory of Intelligent Computing and information processing of the Ministry of education, Hunan Key Laboratory of scientific engineering calculation and numerical simulation, Key Laboratory of engineering structure dynamics and reliability analysis of Hunan Key Laboratory of national defense science and technology and Hunan Provincial Patent Analysis and evaluation center Taiwan. Based on the industrial development of mathematics, statistics and data industry, we will explore the training mode of compound applied science professionals who master modern mathematics and statistics ideas and methods, have deep science foundation and strong engineering application ability, Xiangtan University is a pilot unit to study and formulate the talent training program and curriculum system of "data computing and application" of Applied Science, plan and carry out the construction of series of teaching materials, and promote and lead the "emerging engineering education" derived from science.

6 Mining Subject Words of Data Science Talent Demand

6.1 Overview of LDA Model

In most cases, LDA has two meanings. The LDA topic model in this paper refers to the latent Dirichlet allocation (LDA) model, which was constructed by David BLEI, Andrew ng and Michael I. Jordan in 2003 on the basis of PLSA. It will show the topic of each text in the text set in the form of probability distribution [7]. This is the core definition of generative model, which holds that all words in any text should conform to this rule. Bag of words (BW) is the most common method of LDA model. Each text is regarded as a word frequency vector, which makes it easier to model and analyze after text information is transformed into digital information. LDA model is usually composed of document (d), topic (z) and word (W), so it is also called three-layer Bayesian probability model, assuming that any document is a mixture of various topics, The model is generated from Dirichlet distribution by sampling, and the specific model flow is shown in Fig. 3. Taking the probability information into account in the original traditional space vector model can not only mine the topics in the data set, but also help to extract the hot concerns and related feature words in the data set for in-depth analysis.

6.2 Topic Number Selection

Firstly, input the data processed by feature, and then use M algorithm to solve the topic model to get the corresponding distribution of document, topic and word. The selection of topic number k is the most critical step in setting model parameters. If the value of K is too small, it will make many concepts can not be classified under the corresponding related topics, so they are wrongly summarized to the topics that are not suitable for them, and there is no way to accurately express the concepts that the document needs to convey, which is not conducive to extending the model to other talent demand mining. In

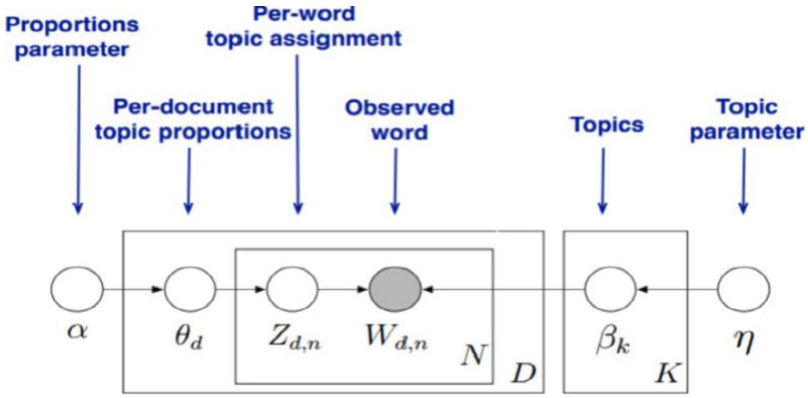

Fig. 3. Flow chart of LDA model

order to help more talent demand data to match with the theme model, the coarse-grained topics such as personal ability and educational background are selected as the theme of the theme model [8].

When the number of topics is set to 3, it means that only three different topics are extracted for analysis. In Fig. 4, we can see that topic 1 has a large number of words, including "data analysis", "algorithm", "machine learning", "data mining", "experience" and other words about professional knowledge, skills and work experience requirements of data science [9]. Topic 2 is mainly about the ability and quality of job seekers, such as "communication", "cooperation", "teamwork" and so on. Topic 3 is

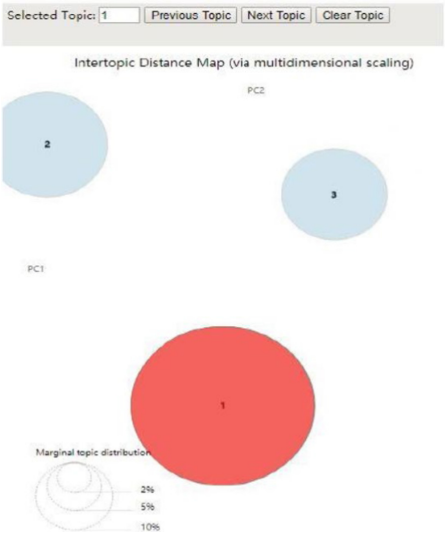

Fig. 4. Topic model visualization when K = 3

about the distribution of academic background words, such as "Statistics", "Mathematics", "postgraduate", etc. [10]. So when the model selects three topics, the result is that the three topics are professional knowledge and skills, personal ability and quality, and education background. From the above analysis of the distribution of words, the result is not ideal when the number of topics is selected as 3. The three topics are divided into multiple topics, so that it is difficult to determine the center of the topic.

6.3 Subject Term Extraction of Data Science Talent Demand

Because the number of topics selected in the best model is four topics, we need to manually delete some words that are not related to the topic content or have less information [11]. Finally, we extract ten most representative topic words from the original topic words. The number of topics selected in this paper is 4, and the corresponding topics are education background, work experience, personal ability and professional knowledge and skills of data science talents. Through the analysis of the keywords of the four categories under the demand of all data science talents, it can be seen that the requirements of professional knowledge and skills are rich and diverse, the work experience and education background are slightly different, and the personal ability and quality are basically the same [12].

In order to highlight the diversity needs of professional knowledge and skills, four types of representative talents are selected for comparative analysis [13]. For the analysis of professional knowledge and skills of data science talents: data analysis talents are mainly required to have good statistical foundation and analysis ability in the application of big data; the professional knowledge and skills of database R & D talents lie in database related knowledge, including database development and operation and maintenance; Algorithm engineering talents mainly need to master the knowledge of machine learning and natural language processing; artificial intelligence development talents need to be familiar with the content of computer and software development and other fields [14].

According to the comprehensive analysis of the other three topics of data science talent demand, the personal ability and quality requirements of the four types of talents are related to teamwork, sense of responsibility, logical thinking and understanding ability; for education background, the subject words of the four types of talents are distributed in the same sentence, It is mainly reflected in the "Statistics", "Mathematics" and "computer" professional "undergraduate" and "graduate" crowd; work experience is similar, which requires us to use word2vec algorithm model to adjust the results of LDA model, and further analyze the demand for data science talents. Through the simple analysis of the demand for four types of representative talents, the topic of LDA topic model is still relatively clear, but the setting of quantity and precision is too rough, which can not completely present the main content of the demand. Therefore, we need to use the word vector model to expand the in-depth mining of topic words [15].

6.4 Overview of word2vec Model

Word2vec is an open source tool for computing word vector in natural language processing. It is a word vector model based on neural network. It can be trained efficiently on

millions of dictionaries and hundreds of millions of datasets, and its algorithm mainly includes cbow model and skip gram model. Cbow is the abbreviation of continua bag of words. Mathematically, it is equivalent to multiplying the vector of a bag of words model by an embedding matrix to get a continuous embedding vector. The model learns the expression of word vector from context's prediction of targe word. The essence of skip gram model is to calculate the cosine similarity between the input vector of input word and the output vector of target word, and to normalize it with SoftMan κ. In this paper, we calculate the cosine value between the numerical morpheme vectors which contain a lot of semantic information through the model, and reflect the degree of association between words in the topic through it, so that the words in the bag of words model will not appear large differences in dimension and semantics, which will affect the result analysis. At the same time, the running speed of the model also has certain advantages compared with other models [16].

6.5 Overview of Visualization Tools

Gepi is a JM based open source and cross platform data visualization software in the field of complex network analysis. It simplifies the network transformation into the form of nodes and edges, and uses them to express the internal structure of data and the relationship between various parts. The visualization model analysis provided by Gepi can be divided into two types: one is to present the location of nodes as a graph in a certain way by selecting different layout algorithms, and interpret the network relationship on this basis. The other is to analyze and explain the network relationship by selecting different statistical algorithms to calculate the overall characteristics, modularity and node centrality of the network.

Gepi provides 12 layout modes, including 6 main layout tools and 6 auxiliary layout tools. The most commonly used are: two force oriented algorithms, circular layout and Hu Yifan layout. Force oriented layout is divided into force atlas and force atlas 2. By imitating the gravitational and repulsive forces of the physical world, force oriented layout automatically generates beautiful network layout graphics until the forces are balanced, and fully shows the overall structure and Automorphism characteristics of the network. It has strong readability, so it plays a leading role in the selection of visual model layout.

Circular layout, also known as fruchterman Reingold layout, is a fr algorithm proposed after many improvements on elastic model. The nodes with edge connection should be close to each other, and the distance between nodes should not be too small, which are two indispensable principles of the algorithm. The algorithm assumes that the nodes in the graph are atoms in particle physics, and calculates the position relationship between the nodes in the form of simulating the position law of atoms until they enter the dynamic equilibrium state. The interaction of gravity and repulsion is considered in the calculation, and the theoretical law of particle physics must be followed. Yifan Hu, Yi Fan Hu proportion and VI I fan Hu multi-level layout are collectively referred to as Hu Yifan layout, which is suitable for large graphics processing, characterized by rough graphics, reducing the amount of calculation and improving the running speed.

Based on LDA topic model and word2wvec word vector model, this section extracts the expanded topic word set, and uses Gepi software to analyze the topic words of data

science talent demand through network relationship analysis method. There are four main themes of data science talent demand: education background, work experience, professional knowledge and skills and personal ability and quality. Based on these four different demand topics, the demand of data science talents is visualized and analyzed, and the core demand of data science talents, the correlation between the internal keywords of each demand topic and the community relationship between each demand topic are mined.

7 Conclusions

On the basis of optimizing the teachers' team in the school, we should build a double qualified team, formulate appropriate teaching resource construction plan, and provide rich teaching resources for big data major through school enterprise cooperation based on the concept of "construction, application and sharing"; we should break down the barriers between colleges and departments and conduct research in the way of interdisciplinary integration; actively carry out multi-channel and multi-mode practice teaching relying on practice base, The experimental teaching mode of "foundation, synthesis and innovation" is constructed. Finally, a complete set of talent training mode is formed, which covers the research and development of big data and industrial application, reflects the school running characteristics of our university and the characteristics of our data department and big data technology specialty, and plays a demonstration and promotion role in the national local colleges and universities.

Acknowledgements. The key Research topics for research on the reform of higher education of Jilin Province Education Department (SJZD19–04) .

References

1. Lu, L., Tian, Z., Zhou, M.: Fast frequency domain synchronization algorithm of main synchronous signal in TD-LTE system. Sci. Technol. Eng. **16**(10), 174–177 (2016)
2. Zhang, P.: Analysis on the cultivation platform of innovative talents in Colleges and Universities Based on 5g technology--Taking the construction of Applied Innovation Laboratory in Guangzhou as an example, human resources development, (9) (2020)
3. Chen, X., Zhou, L., Cao, Y.: Exploring the construction of talent training program for application-oriented Undergraduate data science and big data technology. Mod. Ind. Econ. Inf. **7**(23), 40–42 (2017)
4. Xinyou, L., Ge, L.: Research on big data talent cultivation in higher vocational colleges. J. Hebei Tourism Vocat. Coll. **22**(01), 88–90 (2017)
5. Ji, X.: Research and application of multi label text classification algorithm. Shandong University (2019)
6. Zhu, X.: Microblog recommendation based on wrd2vec topic extraction. Beijing University of Technology (2014)
7. Bonhard, P., Sasse, M.A.: Knowing me, knowing you' — Using profiles and social networking to improve recommender systems. BT Technol. J. **24**, 84–98 (2006). https://doi.org/10.1007/s10550-006-0080-3

8. Chen, X.: Research on some key technologies of text mining. Fudan University (2005)
9. Li, R.: Research on text classification and related technologies. Fudan University (2005)
10. Chen, Z., He, T.: Demand and cultivation of data science talents. Big Data, **2**(05), 95 (2016)
11. Cao, G., Hu, Z., Guo, J., Wang, Y.: Research on professional training requirements and curriculum of master of data science in the United States. Digit. Libr. Forum, (05), 38–45 (2018)
12. Yueliang, Z.: Characteristics and Enlightenment of talent training mode of foreign I schools data science project . Libr. Inf. Knowl. **04**, 109–118 (2018)
13. Qiangshen, W.: Domain keyword extraction: combining LDA with wor d2vec. Guizhou Normal University (2016)
14. Wang, J., Zhang, J.: Application of statistical model in Chinese text mining. Math. Stat. Manage. **36**(04), 609–619 (2017)
15. Shuyuan, N.: Research on the development of data science and personnel training. Stat. Inf. Forum **34**, 117–122 (2019)
16. Zhang, K., Shi, T., Li, W., Qian, R.: Research on wrd2vec optimization strategy based on statistical language model. Chin. J. Inf. **33**(07), 11–19 (2019)